DOCUMENT RESUME

ED 173 381 TM 009 010

AUTHOR Huberty, Carl J.; Mourad, Salah A.
TITLE Variable Selection in Discriminant Analysis.
PUB DATE Apr 79
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Discriminant Analysis; Evaluation Methods; Group Membership; *Groups; Higher Education; Prediction; *Predictor Variables; Research Design; Sampling

ABSTRACT

Methods for ordering and selecting variables for discriminant analysis in multiple group comparison or group prediction studies include: univariate Fs, stepwise analysis, learning discriminant function (LDF) variable correlations, communalities, LDF standardized coefficients, and weighted standardized coefficients. Five indices based on distance, association, and classification accuracy, evaluate the sense of discriminatory power in which a subset of a given size determined by one method, would be better than a subset of the same size determined by another method. To illustrate the use of the methods and indexes, thirteen variables for assigning 153 college freshmen to one of three French courses were rank ordered according to all six methods. The five indexes were computed for subsets of each of ten sizes determined by each method. The relative effectiveness of the variable selection methods for yielding good subsets depended on the index used. The methods agreed closely for the distance and association indexes for all subset sizes (concordance coefficient .91); but they were inconsistent for the classification index (concordance coefficient .17). The combination of variable selection method and purpose of discriminant analysis research, and the relationship between selection method and structure of data need further study. (CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED173381

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Variable Selection in Discriminant Analysis

Carl J Huberty Salah A. Mourad

University of Georgia

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. J. Huberty

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

TM009 010

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979

Abstract

The four purposes of this paper are: (a) To discuss some methods (univariate Fs, stepwise, LDF-variable correlations, communalities, LDF standardized coefficients, weighted standardized coefficients) for ordering and selecting variables in discriminant analysis; (b) To discuss some indices (based on distance, association, and classification accuracy) to use in comparing the relative effectiveness of the methods; (c) To illustrate the use and results of the methods and indices with real data; (d) To make suggestions regarding considerations for more detailed study of the problem.

Variable Selection in Discriminant Analysis

Two aspects of an interpretation problem in multivariate analyses deal with (a) variable ordering, and (b) variable selection. This interpretation problem appears in single group studies (e.g., multiple and canonical correlation analysis) and in multiple group studies (e.g., discriminant analysis). The current focus is on the interpretation problem in discriminant analysis. Discriminant analysis may be considered applicable in group comparison studies or in multiple-group prediction studies. In group comparison studies interest is on obtaining an ordering of the response variables with respect to their relative contribution to group separation, or on obtaining a subset of a given size of the response variables that separates the group mean vectors to a degree greater than that yielded by other subsets of the same size. In prediction studies interest is on determining an ordering or subset of response variables related to accuracy of prediction, i.e., to group membership classification accuracy.

The problem of variable ordering/selection has been studied in the two-group situation (Cochran, 1964; Eisenbeis, Gilbert, & Avery, 1973), and in the more general K-group situation (Eisenbeis & Avery, 1972; Henschke & Chen, 1974; Huberty, 1975).

The four purposes of the current paper are: (a) to discuss some methods for ordering and selecting variables; (b) to discuss some indices to use in

comparing the relative effectiveness of the methods; (c) to illustrate the use and results of the methods and indices with real data; (d) to make suggestions regarding considerations for more detail study of the variable ordering/selection problem.

Variable Ordering/Selection Methods

In a p-variable normal, K-group situation, numerous methods have been proposed for rank-ordering the p variables in terms of their relative "discriminatory power." These rank-orderings may then be used to select variable subsets of size q from a given set of p variables. With a small number of variables, it may be possible to consider all possible combinations of variables for a subset of size q. Although theoretically this may be considered the ideal, it is not practical in many situations. More practical methods have been advanced.

One method considered here involves ordering the p univariate F values. With this method (Method I), a selected subset of size q consists of variables having the q highest F values. A second method is based on the results of a forward stepwise discriminant analysis computer algorithm. With this method (Method II) the first q variables entered into the analysis constitute a "best" subset of size q. A third method (Method III) involves correlation coefficients between each of the p variables and the first (or, leading) linear discriminant function (LDF), with a selected subset of size q consisting of the variables having the q highest squared LDF-variable correlation coefficients. Method IV involves correlations as in the immediately preceding method except that all of the LDFs are considered. For a given variable the LDF-variable correlations are squared and summed across the LDFs; the "best" subset of size q consists of those variables with the q highest sums (called "communalities").

The last two methods under consideration involve the LDF coefficients applicable to standardized variable scores. Method V uses the standardized coefficients of the first LDF, with a subset of size q consisting of the variables with the q largest coefficients (in absolute values). For Method VI the standardized coefficients for each LDF are weighted by the corresponding eigenvalue, with a subset of size q consisting of the variables with the q largest (in absolute value) sums of eigenvalue-coefficient products (see Eisenbeis & Avery, 1972, p. 70).

Discriminatory Indices

In assessing the relative goodness of these methods it is necessary to consider the sense of discriminatory power in which a subset of a given size determined by one method would be better than a subset of the same size determined by another method. Three senses are currently considered. One subset may be judged better than another subset if it yields more spread among the K group mean vectors. The two spread indices used here involve Mahalanobis-type distances: (a) Rao's (1952, p. 257) V statistic, which is a weighted (by the group sizes) sum of the p-variate distances between each group mean vector and the combined-group mean vector, and (b) the Mahalanobis p-variate distance, D, between the closest pair of groups. A second sense of discriminatory power of a variable subset pertains to p-variate strength of association. The two association indices used here are (a) a multivariate generalization of eta squared, η^2_{mult} (Cooley & Lohnes, 1971, p.227), and (b) a multivariate generalization of omega squared, ω^2_{mult} (Tatsuoka, 1970, p.48). The third sense of discrimination power of a

variable subset is reflected in the proportion (P) of cases that is correctly classified. In sum, two inter-group distance indices are considered (V and D), along with two strength of association indices (η^2_{mult} and ω^2_{mult}), and one index of predictive accuracy (P).

Illustrative Example

The data set used to illustrate the application of the methods and indices consists of measures on 13 variables for 153 college freshmen for the purpose of assigning them to one of three French courses. The 13 measures included five high school grade point averages, six standardized achievement test scores, number of semesters of high school French, and number of semesters since last high school French course was taken. Group 1 ($n_1 = 35$) consists of those freshmen assigned to the beginning French course. Groups 2 and 3 ($n_2 = 81$, $n_3 = 37$) consist of freshmen in more advanced French courses. Descriptive information on the data set is given in Table 1.

Insert Table 1 about here

The data analysis started by using all 13 variables. The rank-orderings of the variables according to all six methods were determined (see Table 2).

Insert Table 2 about here

The rankings of the variables yielded by Methods V and VI were identical. This result was due to the fact that the first eigenvalue (2.84) dominated the second eigenvalue (0.13), thus making the sum of eigenvalue-coefficient products heavily dependent on the products for the first LDF. Hence, the discussion of the illustration will involve only Methods I through V. The rank-orderings of all 13 variables across the five methods exhibited some variability; the resulting coefficient of concordance was 0.66.

Subsets of size 1, 2, ..., 12 were specified according to the first five methods. (The condition of equal covariance matrices was considered tenable for each subset.) The five indices of discriminatory power were computed for subsets of each size determined by each method.

For the illustrative data, variable subsets of size 2 yielded values of each of the five indices of discriminatory power that were very nearly the same across all five ordering methods. For subsets of size 1, all methods except Method IV (communalities) were in complete agreement for all five indices of discriminatory power. Method IV was clearly inferior to the other four methods, in all three senses of discriminatory power (i.e., for all five indices), in determining the "best" single variable.

The discussion, then, will focus on subsets of sizes 3 through 12. To give the reader an idea of the obtained values of the five indices, values

are given in Table 3 for subsets of sizes 4 and 9.

Insert Table 3 about here

For subsets of sizes 3 through 12, the five methods were in fairly close agreement for two senses of discriminatory power, distance (indices V and D) and association (indices η^2_{mult} and ω^2_{mult}). The methods were rank-ordered with respect to each index, and a coefficient of concordance was computed for each subset size. The median coefficient across the ten subset sizes was 0.91. Thus, the four indices V, D, η^2_{mult}, and ω^2_{mult}, indicated nearly the same relative effectiveness of the five methods. For each subset size a consensus ordering of the methods was determined. The "average" ordering of the methods over the ten subset sizes was: II (stepwise), IV (communalities), III (first LDF-variable correlations), I (univariate Fs), and V (standardized coefficients). The very same conclusion was reached when the actual numerical index values were summed over the ten subset sizes, followed by an examination of concordance of the four rankings of the five methods based on the sums; the resulting coefficient of concordance was 0.96.

The relative effectiveness of the five methods as assessed by the classification index, P, was fairly inconsistent across the ten subset sizes: coefficient of concordance was .17. However, an examination of both the resulting proportions and rankings of the proportions revealed that Methods II, III,

and IV were approximately equally effective, and clearly more effective than Methods I and V. The "average" ordering of the methods across the ten subset sizes was: III, II, IV, V, and I.

Over the ten subset sizes the relative effectiveness of the five methods varied somewhat from that based on the distance and association indices. For each sample size a Spearman coefficient was computed between the method ranks based on P and the consensus method ranks based on V, D, η^2_{mult}, and ω^2_{mult}. The median coefficient across the ten subset sizes was 0.66.

Comments

To draw any general conclusions regarding the relative effectiveness of the five variable selection methods on the basis of the above results would be folly, indeed. The intent of analyzing these data was simply to present some alternative variable selection methods, and to point out the availability of different indices of discrimination power to assess effectiveness of variable selection methods.

For the data studied here the relative effectiveness of the five methods for yielding "good" subsets appeared to depend on the sense of discriminatory power considered. Whether there is a clear dependence when other data sets are involved is, of course, a question that needs further study. That is, one question needing extended study pertains to the combination of variable

selection method and purpose of research (group comparison or group membership prediction). A second question of interest concerns the relationship between selection method and structure of data. By data structure is meant (1) the degree of multicollinearity of (or, high correlations among) the response variables, (2) the number of "significant" eigenvalues, and (3) the degree of group mean vector separation and mean vector configuration. Another question deals with the relationship among selection method, discriminatory power index, and discrimination among group subsets. The latter is in contrast to separation, association, or classification accuracy across all K groups in an omnibus manner.

The five methods and five indices suggested in this paper do not take the problem of sampling fluctuation into full consideration. In one sense this problem is related to one aspect of data structure; high degrees of collinearity among the response variables lead to unstable standardized weights, and thus may render Methods V and VI inappropriate. One possibility for further study would be to incorporate "ridge" estimates of LDF weights. The incorporation of "adjusted" indices of discriminatory power might also be considered; however, adjustments should not be overstressed for selection purposes,because in selection only the comparison of index values is of primary importance.

We do not pretend to exhaust all selection methods with the five presented. Many others have been proffered along with some comparisons (e.g., Farmer & Freund, 1975). Further, more involved computational algorithms and criteria have been proposed (e.g., McCabe, 1975; Habbema & Hermans, 1977).

It should be noted that even though one selection method may indicate a subset of variables of a given size that is better, in some sense, than that indicated by any other method, that subset may not be the best of all possible subsets of that size. So, one method may be better than all others studied but still not very good in an absolute sense. Of course, one could examine all possible subsets of a given size, but, even in the computerized world, this is very expensive.

For the special case involving only K = 2 groups, choosing a method and an index is simplified somewhat. Here, Methods I and III (and, of course, IV) are equivalent (Huberty, 1972). Also, indices D and V would yield equivalent results in the two-group case.

One last comment about variable selection. If the purpose of a study is that of predicting group membership (when a "classification" analysis is appropriate), deletion of variables must be done with caution. An exploration of the effects of variable deletion must be made. The criterion for retaining or deleting variables should be related to the efficiency of the classification results, both in terms of the overall proportion of correct classification and the separate group proportions.

Table 1

Means, Standard Deviations[a], Univariate Fs, and Within-Groups Correlation Coefficients

Variable	Group 1 (n_1 = 35)	Group 2 (n_2 = 81)	Group 3 (n_3 = 37)	F	Variable 2	3	4	5	6	7	8	9	10	11	12	13
1	2.91 (0.50)	3.28 (0.51)	3.33 (0.49)	8.50	.45	.60	.35	-.29	.60	.37	.15	.26	.18	-.21	.05	-.05
2	2.52 (0.70)	2.86 (0.80)	2.69 (0.80)	2.40		.43	.67	-.21	.55	.28	.53	.26	.32	-.21	-.05	.15
3	2.89 (0.48)	3.34 (0.52)	3.21 (0.55)	9.27			.59	-.20	.51	.05	.18	.22	.12	-.24	-.11	-.02
4	2.65 (0.54)	3.07 (0.60)	2.95 (0.69)	5.84				-.03	.51	.23	.37	.32	.29	-.35	-.15	.03
5	3.17 (0.98)	4.24 (0.76)	4.97 (1.11)	35.60					-.35	-.26	-.18	.02	-.04	.29	.26	.03
6	2.68 (0.70)	3.20 (0.64)	3.15 (0.67)	8.23						.39	.32	.23	.20	-.07	.02	-.16
7	22.43 (2.09)	23.80 (2.86)	25.03 (2.51)	8.84							.36	.33	.40	.07	.05	.04
8	23.66 (5.73)	24.79 (4.99)	24.27 (5.35)	0.58								.38	.53	-.03	-.07	.24
9	25.14 (3.96)	26.59 (3.14)	27.32 (3.68)	3.77									.59	.13	.14	.06
10	23.31 (4.16)	24.63 (4.57)	26.05 (4.61)	3.36										.20	.06	.15
11	14.34 (7.55)	19.27 (7.21)	25.27 (4.01)	67.87											.21	-.16
12	40.86 (7.55)	52.88 (7.21)	63.76 (4.01)	105.90												-.35
13	2.00 (2.16)	1.40 (1.60)	0.49 (0.99)	7.97												

[a]Standard deviations are given in parentheses.

Table 2

Rankings of Variables

Variable	I (F's)	II (Stepwise)	III (LDF-Var r's)	IV (Comm)	V (Stan Coeff)	VI (Wtd Coeff)
1	6	13	7	12	6	6
2	12	10	6	6	12	12
3	4	5	1	8	8	8
4	9	3	4	3	11	11
5	3	4	8	4	3	3
6	7	8	5	13	7	7
7	5	7	10	7	4	4
8	13	12	11	11	13	13
9	10	6	12	5	9	9
10	11	11	13	10	10	10
11	2	2	3	2	2	2
12	1	1	2	1	1	1
13	8	9	9	9	5	5

Table 3

Values for Discriminatory Indices

for Subsets of Sizes 4 and 9

Subset Size	Index	Method I	II	III	IV	V
4	V	354	379	365	379	313
	D	9.77	10.5	10.1	10.5	8.69
	η^2_{mult}	.720	.729	.724	.729	.681
	ω^2_{mult}	.715	.724	.719	.724	.675
	P	850	.863	.850	.863	.810
9	V	402	429	407	435	391
	D	11.0	11.8	11.2	12.0	10.8
	η^2_{mult}	.749	.761	.751	.761	.744
	ω^2_{mult}	.744	.756	.747	.756	.739
	P	.850	.850	.879	.863	.863

REFERENCES

Cochran, W. G. On the performance of the linear discriminant function. Technometrics, 1964, 6, 179-190.

Cooley, W. W., & Lohnes, P. R. Multivariate data analysis. New York: Wiley, 1971.

Eisenbeis, R. A., & Avery, R. B. Discriminant analysis and classification procedures. Lexington, Mass: Heath, 1972.

Eisenbeis, R. A., Gilbert, G. G., & Avery, R. B. Investigating the relative importance of individual variables and variable subsets in discriminant analysis. Communications in Statistics, 1973, 2, 205-219.

Farmer, J. H., & Freund, R. J. Variable selection in the multivariate analysis of variance (MANOVA). Communications in Statistics, 1975, 4, 87-98.

Habbema, J. D. F., & Hermans, J. Selection of variables in discriminant analysis by F-statistic and error rate. Technometrics, 1977, 19, 487-493.

Henschke, C. I., & Chen, M. M. Variable selection technique for classification problems. Educational and Psychological Measurement, 1974, 34, 11-18.

Huberty, C. J Regression analysis and 2-group discriminant analysis. Journal of Experimental Education, 1972, 41, 39-41.

Huberty C. J The stability of three indices of relative variable contribution in discriminant analysis. Journal of Experimental Education, 1975, 44, 59-64.

McCabe, G. P. Computations for variable selection in discriminant analysis. Technometrics, 1975, 17, 103-109.

Rao, C. R. Advanced statistical methods in biometric research. New York: Wiley, 1952.

Tatsuoka, M. M. Discriminant analysis. Champaign, Ill.: Institute for Personality and Ability Testing, 1970.